A Unified Probabilistic Model for Learning Latent Factors and Their Connectivities from High-Dimensional Data

Monti, Ricardo Pio

AUAI Press
2018-08-06

Monti , R P & Hyvärinen , A 2018 , A Unified Probabilistic Model for Learning Latent Factors and Their Connectivities from High-Dimensional Data . in A Globerson & R Silva (eds) , Uncertainty in Artificial Intelligence : Proceedings of the Thirty-Fourth Conference (2018) . AUAI Press , Oregon , pp. 300-309 , Conference on Uncertainty in Artificial Intelligence , Monteray , California , United States , 06/08/2018 . < http://auai.org/uai2018/proceedings/papers/123.pdf >

http://hdl.handle.net/10138/309831

unspecified
acceptedVersion

A Unified Probabilistic Model for Learning Latent Factors and Their Connectivities from High-Dimensional Data

Ricardo Pio Monti[1] and **Aapo Hyvärinen**[1,2]
[1]Gatsby Computational Neuroscience Unit, University College London, UK
[2]Department of Computer Science and HIIT, University of Helsinki, Finland

Abstract

Connectivity estimation is challenging in the context of high-dimensional data. A useful preprocessing step is to group variables into clusters, however, it is not always clear how to do so from the perspective of connectivity estimation. Another practical challenge is that we may have data from multiple related classes (e.g., multiple subjects or conditions) and wish to incorporate constraints on the similarities across classes. We propose a probabilistic model which simultaneously performs both a grouping of variables (i.e., detecting community structure) and estimation of connectivities between the groups which correspond to latent variables. The model is essentially a factor analysis model where the factors are allowed to have arbitrary correlations, while the factor loading matrix is constrained to express a community structure. The model can be applied on multiple classes so that the connectivities can be different between the classes, while the community structure is the same for all classes. We propose an efficient estimation algorithm based on score matching, and prove the identifiability of the model. Finally, we present an extension to directed (causal) connectivities over latent variables. Simulations and experiments on fMRI data validate the practical utility of the method.

1 INTRODUCTION

Estimating the connectivity structure between observed variables is a fundamental problem in statistics and machine learning. Probabilistic methods are often based on estimation of the covariance matrix or it inverse. A number of estimators have been proposed for both (Dempster, 1972; Ledoit & Wolf, 2003). On a more general level, such undirected connectivity estimation is a special case of modelling the covariance matrix, which is one of the goals of classical dimensionality reduction methods such as factor analysis and principal components analysis (PCA). In contrast, directed connectivity estimation studies the causal dependence structure across variables (Pearl, 2009). In this work we present methods to perform both directed and undirected connectivity estimation of latent variables in the context of high-dimensional data.

An important problem in practice is that many connectivity estimation methods assume we observe, and know how to choose, the variables between which the connectivity is to be estimated. However, in practice we often have very high-dimensional data, and it may not be useful or feasible to estimate the connectivities between all of them. It is important to somehow reduce the number of variables so that the connectivity estimation is feasible, and furthermore, such reduction can greatly facilitate interpretation of the results. It is often useful to perform the dimension reduction so that it can be interpreted as clustering, as in non-negative PCA (Sigg & Buhmann, 2008). A relevant challenge is how such reduction in the number of variables should be combined with connectivity estimation. In the past, approaches based on stochastic block models (Airoldi et al., 2008; Marlin & Murphy, 2009) or clustered factor analysis (Buesing et al., 2014) have been employed. However, such methods do not explicitly model the connectivity over latent variables and cannot easily be extended to accommodate multiple classes of related datasets, both of which are of interest in this work. Alternative methods recover correlation structure over latent variables but do not focus on dimensionality reduction (Sasaki et al., 2017). Conversely, in the context of directed connectivity, the causal clustering of observed variables has been studied by Silva et al. (2006), Shimizu et al. (2009) and Kummerfeld & Ramsey (2016).

A further challenge is estimating multiple related connectivity matrices, assuming the data is divided into a number of classes, such as subjects in a biomedical setting.

While the estimation of multiple related Gaussian graphical models, parameterized by the inverse covariance, has been extensively studied (Varoquaux et al., 2010; Danaher et al., 2014; Monti et al., 2017), we want to combine such multiple connectivity estimation with the variable reduction scheme described above as well as extend such methods to the domain of directed connectivities.

A practical application that motivates our theoretical developments is functional MRI (fMRI) data, where estimation of "functional connectivity" is widely practiced. Such analysis is a cornerstone of modern neuroscientific research, having provided fundamental insights into the structure and architecture of the human connectome. However, the existing methods are often not very rigorous and would benefit from a proper probabilistic formulation. Traditionally, functional connectivity networks have been modeled as covariance graphs, where the nodes in the network correspond to spatially remote brain regions and edges encode the marginal dependence structure (often simply the covariance). In fMRI, we very clearly see the importance of the theoretical points raised above, in the form of the following challenges:

- Inter-subject consistency: Data is often collected across a cohort of subjects. A hallmark of brain networks is their inter-subject consistency; observed patterns in connectivity have been shown to demonstrate reproducible properties across subjects (Damoiseaux et al., 2006). This suggests significant benefits can be obtained by sharing information across subjects in a judicious manner. In fact, some of the most recent developments are based on collecting hundreds or even thousands of subjects' data in a single data base (Di Martino et al., 2014).

- Modularity: Current methods do not actively incorporate domain knowledge relating to brain networks, a prime example of which is their modular structure which suggests that variables can be aggregated into non-overlapping modules or sub-networks (Sporns & Betzel, 2016). We note this property is not unique to brain networks, but also present in many real-world networks (Newman, 2006).

While motivated by fMRI, we note that these two properties are relevant to wide range of applications such as cyber-security, gene expression data and econometrics.

In this work, we propose a probabilistic latent variable model which is able to directly address the aforementioned issues. The proposed model consists of a low dimensional set of latent variables in a factor analytic model. The associated factor loading matrix is shared across classes and constrained to be non-negative and orthonormal, thereby encoding module/community membership along its columns. Thus, the factors are interpreted as the activities in modules or communities.

Importantly, and in contrast to almost all related models, these latent variables or factors have full (i.e., non-diagonal) covariance structure which we term *latent connectivities*, giving the connectivity structure of the non-overlapping modules. The connectivity structure can be different between classes; however, the model can equally well be applied on data from a single class. Thus, we model both the grouping of variables, and the connectivity between the groups in a single probabilistic model, which can be seen as a variant of factor analysis. We argue that such a formulation leads to important benefits from the viewpoint of interpretation and identifiability whilst remaining plausible from an application perspective.

In contrast to classical Gaussian factor analysis, we are able to prove the uniqueness of the solution: the factors and loadings are identifiable like in (non-Gaussian) independent component analysis (Comon, 1994), largely based on non-negativity inherent in the module structure (Paatero & Tapper, 1994; Seung & Lee, 1999; Donoho & Stodden, 2004). We further propose an efficient parameter estimation algorithm based on score matching. Finally, we demonstrate that the proposed model can be extended to modelling directed connectivities between the latent variables. In this context, the factor loading matrix can be seen as a "pure" measurement model of a Bayesian network (Silva et al., 2006) of causal relationships between high-dimensional observations and their latent variables.

The remainder of this manuscript is organized as follows; in Section 2 we present the proposed model in the context of undirected latent variables. Section 3 provides an identifiability analysis for the proposed method. An efficient estimation algorithm based on score matching is presented in Section 4. The proposed method is extended to recover causal structure over latent variables in Section 5. Experimental results are presented in Section 6.

2 LATENT CONNECTIVITIES MODEL

We propose a latent variable model to accurately find modules (communities, clusters) and model their connectivities, possibly across multiple related classes (conditions, subjects). We assume we have access to multivariate data over N distinct classes, but all our results allow for the simple case $N = 1$ as well. For a given class i, we write $X^{(i)} \in \mathbb{R}^p$ to denote the p-dimensional observed random vector. The ith class is associated with a k-dimensional latent vector, $Z^{(i)}$, which is related to observations, $X^{(i)}$, via a loading matrix $W \in \mathbb{R}^{p \times k}$. We note that the loading matrix is shared across all classes and will serve to encode

module memberships across classes.

We start by a model of undirected connectivities in this section. Here, we assume that the data for each class follows a stationary multivariate Gaussian distribution with zero mean and covariance $\Sigma^{(i)} \in \mathbb{R}^{p \times p}$. Both observations and latent variables are taken to follow multivariate Gaussian distributions, such that:

$$Z^{(i)} \sim \mathcal{N}\left(0, G^{(i)}\right) \tag{1}$$

$$X^{(i)}|Z^{(i)} = z^{(i)} \sim \mathcal{N}\left(Wz^{(i)}, v^{(i)}I\right). \tag{2}$$

If $G^{(i)}$ were diagonal, equations (1) and (2) would correspond to the traditional factor analysis or probabilistic PCA models (Tipping & Bishop, 1999). Our model is able to capture low-rank covariance structure via the loading matrix, W, as follows:

$$\Sigma^{(i)} = WG^{(i)}W^T + v^{(i)}I. \tag{3}$$

From equation (3) it follows that the loading matrix W serves to encode reproducible covariance structure which is present across all classes. In this work, we extend the traditional factor analysis model as follows:

- The loading matrix, W, is constrained to be non-negative and orthonormal. This leads to a loading matrix with at most one non-zero entry per row. We may interpret the columns of W as encoding membership to k non-overlapping modules or sub-networks.
- We introduce latent variables with a non-diagonal covariance structure, which we term *latent connectivities*. While the columns of the loading matrix encode module membership, the non-trivial covariance structure over latent variables may be interpreted as modeling marginal dependencies (i.e., connectivity) across distinct modules or sub-networks. Such an interpretation is very natural in many applied settings.

We note that the introduction of marginally dependent latent variables is not possible in the context of traditional factor analysis, since the effects of factor connectivity and factor loadings cannot be distinguished. In fact, an ordinary factor analysis model is non-indentifiable even with uncorrelated factors. However, in combination with the aforementioned constraints on the loading matrix, it is possible to identify the latent connectivities in our model (see next section). We thus argue that our model is able to capture the modular nature of many real-world datasets and, due to its identifiability, yields easily interpretable results. Figure 1 provides an overview of the proposed model in the context of estimating brain connectivity networks.

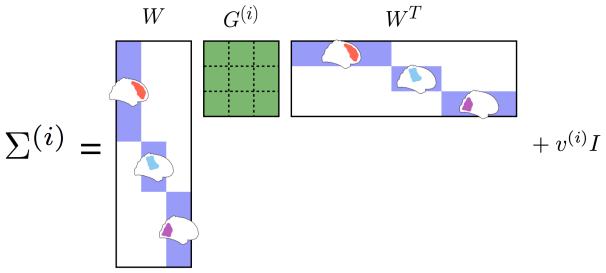

Figure 1: Visualization describing the various components of the proposed covariance model. The factor loading matrix, W, is shared across all subjects and serves to denote membership into non-overlapping brain modules. The *latent connectivity* across modules, parameterized by $G^{(i)}$, is allowed to vary across subjects.

3 IDENTIFIABILITY ANALYSIS

We note that without the introduction of constraints on W, the covariance model proposed in equation (3) is not unique. For example, it would be possible to re-parameterize W such that $G^{(i)}$ is diagonal matrix by using an eigen-value decomposition of $G^{(i)}$. However, the following properties demonstrate that non-negativity and orthonormality constraints are sufficient to ensure the solution is identifiable.

Property 1. *Assume non-negative, orthonormal W. Then at most one entry per row of W can be non-zero.*

Proof. Directly from the constraints on W, we can express the (i, j) entry of W^TW as:

$$\left(W^TW\right)_{ij} = \sum_{r=1}^{k} \left(W^T\right)_{ir} W_{rj} = \sum_{r=1}^{k} W_{ri}W_{rj} = \delta_{ij}$$

which, combined with non-negativity, implies that the i and j columns of W can have no overlapping support for $i \neq j$. □

Property 2. *Assume non-negative, orthonormal W. Then any matrix $V \in \mathbb{R}^{k \times k}$ for which $\tilde{W} = WV$ is non-negative and orthonormal must be the identity matrix or some permutation of the identity.*

Proof. By Property 1 we have that both W and $\tilde{W}$ have at most one non-zero entry per row. Since $\tilde{W}^T\tilde{W} = I$ we have that $V^TV = I$. Define c_i and $\tilde{c}_i$ to be the index of the non-zero entry along the ith row for W and $\tilde{W}$ respectively. By construction:

$$\tilde{W}_{i\tilde{c}_i} = (WV)_{i\tilde{c}_i} = \sum_{r=1}^{k} W_{ir}V_{r\tilde{c}_i} = W_{ic_i}V_{c_i\tilde{c}_i} > 0$$

where the final equality follows from the fact that W_{ic_i} is the only non-zero entry along the ith row of W. Since $W_{ic_i} > 0$, this implies that $V_{c_i\tilde{c}_i} > 0$. Furthermore, for $j \neq \tilde{c}_i$ we have

$$\tilde{W}_{ij} = (WV)_{ij} = \sum_{r=1}^{k} W_{ir}V_{rj} = W_{ic_i}V_{c_ij} = 0$$

Since $W_{ic_i} > 0$, we must have that $V_{c_ij} = 0$ whenever $j \neq \tilde{c}_i$. When combined with the fact that $V^TV = I$, it follows that V must either be the identity matrix of a permutation it. □

Property 2 indicates that the matrix W is uniquely defined in our model, and there is nothing like an undetermined factor rotation in conventional Gaussian factor analysis. By similar logic, Property 2 also implies the uniqueness of $G^{(i)}$.

4 ESTIMATION BY SCORE MATCHING

The parameters associated with the proposed model consist of the loading matrix, W, the latent variable covariances, $\{G^{(i)}\}$, and the observation noise, $\{v^{(i)}\}$. One potential strategy is to estimate latent variables in an expectation-maximization framework. However, due to relative simplicity of the proposed covariance model we propose to directly marginalize out latent variables.

Parameters may also be estimated via maximum likelihood estimation, however, this results in an iterative algorithm where the computational cost of each parameter update is $\mathcal{O}(p^3)$ (a derivation of which is provided in the Supplementary materials). Instead, we propose to estimate parameters by score matching (Hyvärinen, 2005), leading to an algorithm with a computational cost of $\mathcal{O}(p^2k)$ per iteration. This is a significant reduction as we will typically expect $k \ll p$. While score matching is typically used in the context of unnormalized statistical models, it may often result in optimization-related benefits for normalized models as well (Hyvärinen, 2007; Lin et al., 2016).

In the context of multivariate Gaussian data, the score matching objective function is defined as (Hyvärinen, 2005):

$$J = \sum_{i=1}^{N} -\text{tr}\left(\Omega^{(i)}\right) + \frac{1}{2}\text{tr}\left(\Omega^{(i)}\Omega^{(i)}K^{(i)}\right), \qquad (4)$$

where $K^{(i)}$ is the sample covariance matrix for class i and $\Omega^{(i)}$ is the inverse covariance, which may be computed by the Sherman-Woodbury identity as:

$$\Omega^{(i)} = {v^{(i)}}^{-1}\left(I - WG^{(i)}(G^{(i)} + v^{(i)}I)^{-1}W^T\right).$$

Directly optimizing the score matching objective (equation (4)) under non-negativity and orthonormality constraints on the loading matrix is challenging. One potential strategy is to employ projected gradient descent as suggested by Hirayama et al. (2016). However, projecting onto the non-negative Stiefel manifold is undesirable as it requires W to have at most one non-zero entry per row at each step of the optimization algorithm (see Property 1 above). Such an approach is therefore highly dependent to the random initialization of the loading matrix.

In this work we seek to minimize equation (4) in a constrained optimization framework. This allows for the orthonormality constraints to be enforced via an augmented Lagrangian penalty (Bertsekas, 2014), while the non-negativity is enforced at each iteration by projecting onto the non-negative orthant.

The objective function associated with the augmented Lagrangian is defined as:

$$\tilde{J} = J + \frac{\rho}{2}||W^TW - I_k||_2^2 + \text{tr}(\Lambda^T(W^TW - I_k)),$$

where $\Lambda \in \mathbb{R}^{k\times k}$ are Lagrange multipliers enforcing the orthonormality constraints, ρ is a positive scalar parameter parameterizing the augmented penalty term and J is the original score matching objective. We may then proceed to iteratively optimize each of the parameters using gradient descent. In particular, the gradient of the score matching objective with respect to the loading matrix can be computed as:

$$\frac{\partial J}{\partial W} = -\sum_{i=1}^{N} K^{(i)}WA^{(i)}\left(I - \frac{1}{2}A^{(i)}\right),$$

where we define $A^{(i)} = G^{(i)}(G^{(i)} + v^{(i)}I)^{-1}$. Similarly, in the case of the latent connectivities, $G^{(i)}$, and observation noise, $v^{(i)}$, we have

$$\frac{\partial J}{\partial G^{(i)}} = {v^{(i)}}^{-2}\left[v^{(i)}I - W^TK^{(i)}W\left(I - A^{(i)}\right)\right] \times \left[(G^{(i)} + v^{(i)}I)^{-1}\left(I - A^{(i)}\right)\right].$$

$$\frac{\partial J}{\partial v^{(i)}} = \sum_{i=1}^{N} -{v^{(i)}}^{-3}\text{tr}\left(K^{(i)} - v^{(i)}I\right) + {v^{(i)}}^{-3}\text{tr}\left(W^TK^{(i)}WH_1^{(i)}\right) + H_2^{(i)}$$

where $H_1^{(i)} \in \mathbb{R}^{k\times k}$ and $H_2^{(i)} \in \mathbb{R}$ are defined, together with the relevant derivations, in the Supplementary material.

Estimation proceeds by iteratively updating each of the parameters in a gradient descent framework. In the context of the loading matrix the step-size, η, is selected via the

Armijo rule and we project onto the non-negative orthant at each iteration, resulting in an update of the form:

$$W \leftarrow \mathcal{P}_+ \left(W - \eta \left(\frac{\partial J}{\partial W} + \rho(WW^TW - W) + W\Lambda \right) \right)$$

where $\mathcal{P}_+(x) = \max(0, x)$ is the projection onto the non-negative orthant. Moreover, in the case of latent variable connectivities we have:

$$\frac{\partial J}{\partial G^{(i)}} = 0 \iff G^{(i)}(G^{(i)} + v^{(i)}I)^{-1} = I - v^{(i)}(W^T K^{(i)} W)$$

which after some manipulation yields a closed form update for the latent connectivity structure as:

$$G^{(i)} \leftarrow W^T K^{(i)} W - v^{(i)} I \tag{5}$$

By writing $W^T K^{(i)} W = (X^{(i)}W)^T (X^{(i)}W)$ we note that this update has an intuitive interpretation as the covariance across estimated modules. The Lagrange multipliers are updated as (Bertsekas, 2014):

$$\Lambda \leftarrow \Lambda + \rho(W^TW - I)$$

It is important to note that the proposed method only enforces orthonormality on the loading matrix in the limit of convergence. However, the updates provided above are premised on the assumption that W is orthonormal. As such, the aforementioned updates only correspond to approximations in the case where W is non-orthonormal.

Hyper-parameter Tuning In practice, the proposed model requires the selection of a single hyper-parameter, k, which determines the dimensionality of latent variables. We propose to tune k by minimizing the negative log-likelihood over held-out data.

5 EXTENSION TO DIRECTED CONNECTIVITY

In this section we describe a natural extension of the aforementioned model to estimate causal structure (directed connectivity) across latent variables. As in the previous section, we restrict ourselves to linear latent variables models where each observed variable is conditionally dependent on a single latent variable. In the context of Bayesian networks such models are known as *Pure 1-Factor* models (Silva et al., 2006; Kummerfeld & Ramsey, 2016). We note that such an assumption directly corresponds to each row of the loading matrix, W, containing at most one non-zero entry. This is precisely what is enforced by the non-negativity and orthonormality assumptions introduced in this work. As such, restricting the loading matrix in this manner corresponds to a natural and frequently employed assumption when attempting to recover causal structure (Silva et al., 2006).

We follow Shimizu et al. (2006) and study a non-Gaussian variant of Bayesian networks over latent variables. Formally, we assume variables $Z_j^{(i)}, j \in \{1 \dots k\}$ can be arranged in a causal ordering such no later variables causes a variable ahead of it in the order. We denote such an ordering by $k(j)$ and assume that each variable, $Z_j^{(i)}$, is a linear function of earlier variables together with a non-Gaussian disturbance, $e_j^{(i)}$, such that:

$$Z_j^{(i)} = \sum_{k(r)<k(j)} b_{jr}^{(i)} Z_r^{(i)} + e_j^{(i)}. \tag{6}$$

Due to the linear nature of dependencies, we can write equation (6) as follows:

$$Z^{(i)} = B^{(i)} Z^{(i)} + e^{(i)} \tag{7}$$

$$= \left(I - B^{(i)} \right)^{-1} e^{(i)}. \tag{8}$$

The matrix $B^{(i)}$ encodes a Directed Acyclic Graph (DAG), which corresponds to the *structural model* over latent variables. We note that equation (7) corresponds to a Linear, non-Gaussian, acyclic model (LiNGAM; Shimizu et al., 2006). As in the previous section, observed data are subsequently related as follows:

$$X^{(i)} | Z^{(i)} = z^{(i)} \sim \mathcal{N}(Wz^{(i)}, v^{(i)} I) \tag{9}$$

where the loading matrix encodes the *measurement model*.

To date, a wide range of algorithms have been proposed to estimate measurement models. Prominent examples include the `BuildPureClusters` and `FindOneFactorCluster` algorithms. However, such methods cannot easily be extended to the context of multiple related datasets where the underlying structural models are heterogeneous. Moreover, such methods do not scale well to high-dimensional data. In contrast, our method proposed below can easily accommodate such data and is therefore a good candidate to accurately recover the measurement model. Once the measurement model has been estimated, we may proceed to infer the causal structure over latent variables using established methods such as LiNGAM.

We now outline our two-stage procedure to estimate *Pure 1-Factor* latent variable models. First, the score matching algorithm detailed in Section 4 is employed to estimate the measurement model (i.e., the loading matrix W). Given a measurement model, there are a variety of algorithms to estimate the structural moel. In this work we follow Shimizu et al. (2009) and propose to recover causal dependencies over latent variables by applying LiNGAM to projected observations, $\hat{W}^T X^{(i)}$. This step is performed independently for each of the N classes.

We note that the likelihood proposed in Section 2 is misspecified in the context of non-Gaussian Bayesian networks considered here (as latent variables follow non-Gaussian distribution). However, in the first stage we are only interested in the estimation of the loading matrix, W, using covariance information which is unaffected by non-Gaussianity over latent variables. In fact, as we allow for arbitrary (i.e., non-diagonal) latent connectivities, the proposed model is able to accommodate covariances induced by the causal structure whilst estimating the loading matrix. Alternative approaches, such as the `BuildPureClusters` algorithm, are also based exclusively on studying covariance structure, albeit while introducing additional higher-order algebraic constraints.

6 EXPERIMENTAL RESULTS

We systematically assess the performance of the proposed model using simulated data under the Gaussian factor analysis model of Section 2 as well as the directed latent Pure Bayesian network model described in Section 5. Finally, we present an application to resting-state fMRI data from the ABIDE consortium (Di Martino et al., 2014).

6.1 PERFORMANCE METRICS

We assess the performance of the proposed method in the context of both a single class and multiple related classes. Throughout these simulations, we quantify the performance of various methods based on three distinct tasks:

1. Recovery of the loading matrix, W. This corresponds to accurately recovering the mixing matrix, and by implication, the module memberships for each variable.

2. Recovery of the latent connectivity structure. In the context of undirected latent connectivities, this corresponds to accurately recovering the covariance structure, $G^{(i)}$, across estimated modules. Conversely, in the context of directed latent Bayesian network models it corresponds to accurately recovering the causal dependence structure over latent variables, encoded in $B^{(i)}$.

3. In the context of undirected connectivities we also measure the negative log-likelihood over unseen data. This provides an objective quantification of how well the proposed method is able to model the data in comparison to alternative methods.

6.2 GAUSSIAN FACTOR ANALYSIS MODEL

Data was generated according to the model described in Section 2. The covariance structure for latent variables, $G^{(i)}$, was randomly generated for each class by sampling the lower triangular entries from a standard Gaussian distribution and multiplying by its transpose. We note that generating $G^{(i)}$ in this fashion will randomly introduce both positive and negative correlations across modules. A random loading matrix, W, was generated by sampling uniform random variables and projecting onto the non-negative Stiefel manifold (this involved retaining the largest entry per row and setting all other entries to zero). Observations for each class were subsequently generated according to equations (1) and (2). The dimensionality of observations and latent variables was set to $p = 50$ and $k = 5$ respectively. Data was generated in this manner for N subjects. We consider two cases: $N = 1$ and $N = 10$, which correspond to the single class and multiple class scenarios. Each experiment was repeated 500 times.

The proposed method was benchmarked against several widely used alternatives. In the context of recovering the loading matrix, W, and covariance structure of latent variables, $G^{(i)}$, we compare to non-negative PCA (using the method proposed by Sigg & Buhmann (2008)) and traditional factor analysis (where Varimax rotation was employed, since it should be able to recover module structure in the factor loadings). We note that while these methods do not explicitly model the covariance structure across latent variables, this can be estimated by first projecting observations using the estimated loading matrix and subsequently studying the covariance structure. Indeed, this is related to the update performed by the proposed method in equation (5). When measuring the negative log-likelihood over unseen data we add additional comparisons against the sample covariance matrix and the estimate proposed by Ledoit & Wolf (2003) and the graphical Lasso.

Simulation results for a single ($N = 1$) class are shown along the top panel of Figure 2. The top left panel plots the squared error when estimating the loading matrix. In the presence of small sample sizes, the proposed method is comparable to non-negative PCA. However, as the sample size increases the proposed method consistently outperforms alternative methods as it is able to model the connectivity of latent variables. Additional results relating the clustering implied by estimated loading matrices are provided in the supplementary materials. Results for the estimation of latent connectivities, $G^{(i)}$, are shown in the top middle panel. We note that the proposed method comfortably outperforms competing methods. Finally, the right panel shows mean negative log-likelihood over unseen data; the proposed method consistently out-performs

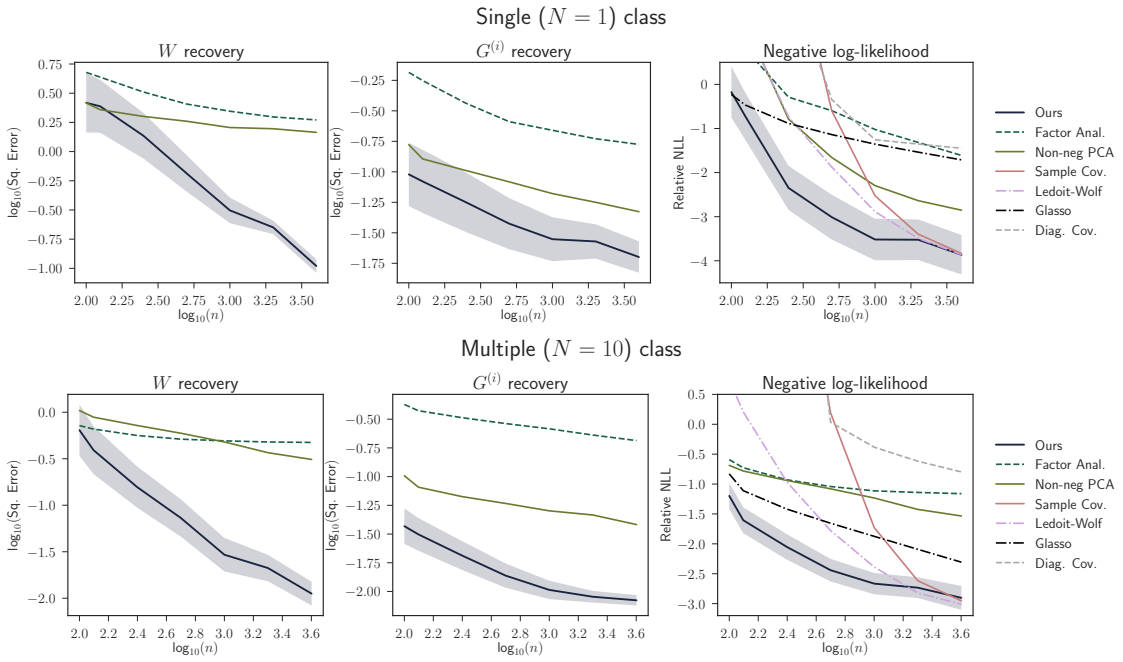

Figure 2: Simulated data results for Gaussian latent variable models with single ($N = 1$) and multiple ($N = 10$) classes are shown along the top and bottom panels respectively. Left and middle panels plot the mean squared error for the estimated loading and latent variable covariance matrices as a function of sample size, n. Right panels shows the mean negative log-likelihood for unseen data as a function of sample size, n. Shaded regions correspond to 95% error bars.

alternative methods for small and moderate sample sizes and remains competitive as sample size increases.

The bottom panel of Figure 2 plots results for the general case of multiple subjects. While the loading matrix is shared across all classes, the latent covariance structure is heterogeneous. While the proposed method is well-suited to accommodate such data, methods such as PCA and factor analysis are not directly applicable. As such, non-negative PCA and factor analysis models were applied using a naive aggregation of the data which concatenated observations across all classes. As before, by accurately modeling the marginal dependencies across latent variables, the proposed method is able to obtain far more accurate estimates of both the loading matrix as well as the latent connectivity structure. The bottom right panel of Figure 2 plots the mean negative log-likelihoods over unseen data and provides empirical evidence that the proposed model provides an accurate estimate of covariance structure.

In addition to quantifying the recovery of the loading matrix and latent connectivities, we also quantify the computation cost associated with each algorithm. The top panel of Figure 4 shows the mean running time as a function of the number of observed variables, p. The results validate our prior claims that the score matching algorithm yields significant computational improvements compared to the maximum likelihood algorithm described in the Supplementary material. For high-dimensional data, the proposed score matching algorithm also improves on the running time compared to both factor analysis and non-negative PCA.

6.3 LATENT PURE BAYESIAN NETWORKS

In this section we perform experiments where the data is generated as described in Section 5. This involved generating latent variables following a non-Gaussian variant of Bayesian networks where the disturbances, $e^{(i)}$, were simulated according to a Logistic distribution. The weights for the structural model, encoded in $B^{(i)}$, were randomly generated together with a distinct random causal ordering for each class. The loading matrix, W, was generated as in Section 6.2.

In the context of recovering the measurement model (i.e., the loading matrix W) we benchmark the proposed method with factor analysis and non-negative PCA as well as the `FindOneFactorClusters` (FOFC) algorithm[1] proposed by Kummerfeld & Ramsey (2016). Formally, the FOFC algorithm only returns non-overlapping

[1]The Tetrad project implementation was employed.

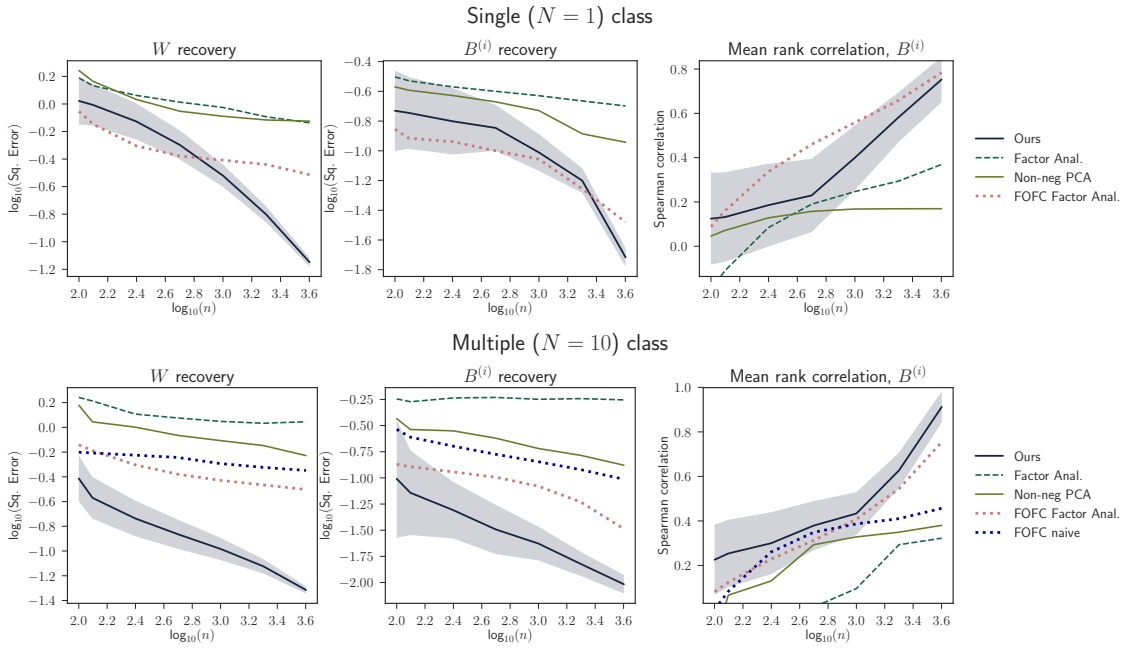

Figure 3: Simulated data results for latent Bayesian networks with single ($N = 1$) and multiple ($N = 10$) classes are shown along the top and bottom panels respectively. Left and middle panels plot the mean squared error for the estimated loading matrix and structural dependency matrices as a function of sample size, n. The right panels show the mean correlation between the estimated causal ordering of latent variables and the true causal order. For all algorithms the causal structure over latent variables was estimated by LiNGAM. Shaded regions correspond to 95% error bars.

clusters of observed variables which share the same latent parent. In order to estimate the associated loading matrix we subsequently employ factor analysis whilst preserving the 1-Factor structure as suggested by Shimizu et al. (2009). Given an accurate estimate of W, we may directly project observations, $\hat{Z}^{(i)} = \hat{W}^T X^{(i)}$ and apply traditional causal discovery algorithms by treating $\hat{Z}^{(i)}$ as observed variables (Silva et al., 2006; Shimizu et al., 2009). Throughout these experiments, the causal structure of latent variables was inferred using LiNGAM.

Figure 3 shows results for a single ($N = 1$) and multiple ($N = 10$) class cases along the top and bottom row respectively. The left panels show the squared error when estimating the loading matrix. In the context of causal models this corresponds to accurately recovering the measurement model. When data is only available for a single class (top left panel) the performance of the proposed method is similar to that of the FOFC algoritm. However, when data across multiple classes is available, the proposed method is able to exploit this information and improve upon the FOFC algorithm as shown in the bottom left panel. We also plot the performance of running the FOFC when naively aggregating data across multiple subjects. Such a naive aggregation leads to worse performance as each class has its own latent causal structure.

We observe a similar pattern when studying the recovery of the structural equations, as shown in the middle and right panels. The proposed method out-performs both factor analysis and PCA and is comparable to FOFC in the context of a single ($N = 1$) class. However, in the context of multiple classes the proposed method is able to out-perform alternative methods.

Finally, the bottom panel of Figure 4 plots the mean running time as the number of observed variables, p, increases. In terms of running time, the proposed method is significantly faster than the FOFC algorithm.

6.4 APPLICATION TO FMRI DATA

In this section we apply the proposed method to resting-state fMRI data taken from the ABIDE consortium (Di Martino et al., 2014). Data was collected from the University of Maryland site corresponding to 53 healthy controls as well as 53 age matched Autism Spectrum Disorder (ASD) subjects. Data from each subject was treated as a distinct class, resulting in $N = 106$ classes. Data were preprocessed via the CPAC pipeline from the ABIDE repository[2]. Time courses were then extracted from 116

[2]http://preprocessed-connectomes-project.org/abide/

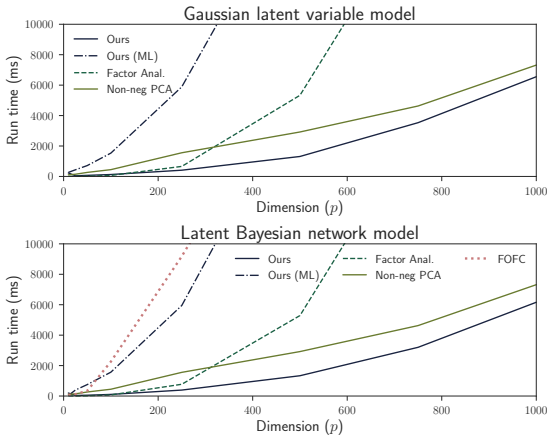

Figure 4: Mean running times (in milliseconds) taken to estimate the factor loading matrix, W, when data is generated according to the Gaussian latent variable model (top) and latent Bayesian network model (bottom). Mean run times based on 10 experiments run on a Macbook Pro (3.5 GHz Intel Core i7, 16 GB RAM).

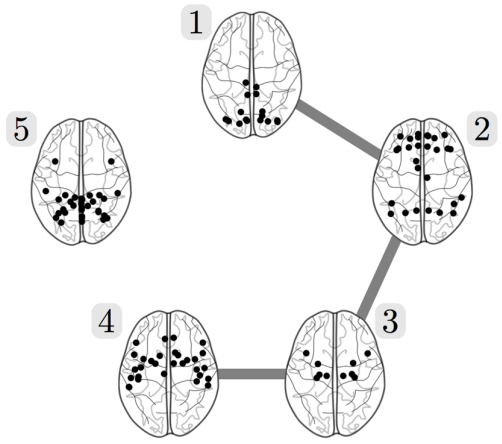

Figure 5: Visualization of estimation brain modules. Each black node represents a distinct brain region. Note that estimated modules are both spatially consistent and symmetric across hemispheres. Edges indicate significant increase in inter-module marginal dependence for ASD subjects compared to healthy controls (edge-wise Bonferroni corrected permutation tests, $p < 0.01$).

regions defined by the Automated Anatomical Labeling (AAL) atlas, yielding 296 observations over 116 nodes for each subject. The data was analyzed under the assumption of undirected latent connectivity structure as there is a large literature discussing differences in covariance structure between healthy controls and ASD subjects (Fox & Greicius, 2010).

The proposed method requires the specification of a single parameter, k, which dictates the dimensionality of latent variables. As discussed in Section 4, this parameter was selected by minimizing the negative log-likelihood over held-out data, resulting in an choice of $k = 5$ modules. Figure 5 shows the $k = 5$ estimated modules obtained by applying the proposed method. The spatially consistency and inter-hemispheric symmetry of module assignments reflects the anatomical and functional architecture of the brain. Moreover, edges in Figure 5 highlight significant differences in covariance structure of latent variables between healthy controls and ASD subjects. Permutation tests were performed on each edge of the latent connectivity structure and Bonferroni corrected for multiple testing. Results indicate that ASD subjects demonstrate increased connectivity for which there is growing evidence (Keown et al., 2013). In particular, we note increased connectivity between the frontoparietal regions (module 2) and both the occipital regions (module 1) and the hippocampus, amydala and temporal lobes (module 3). Finally, Table 1 reports the mean log-likelihood scores on unseen data for all $N = 106$ subjects, demonstrating the proposed model accurately captures covariance structure of fMRI data.

Table 1: Mean log-likelihood scores on unseen data for $N = 106$ subjects (standard deviations are provided in brackets).

Method	Log-likelihood
Ours	**-163.76 (8.86)**
Non-neg. PCA	-190.47 (9.26)
Factor Anal.	-193.89 (8.05)
Glasso	-198.58 (9.05)
Lediot-Wolf	-247.91 (10.88)
Sample Cov.	-329.75 (14.98)

7 CONCLUSION

We have proposed a probabilistic model which simultaneously performs grouping of variables as well as estimation of the *latent connectivities* between groups. The proposed method can be seen as an extension of traditional factor analysis with the important difference that latent variables are allowed to have a full (i.e., non-diagonal) covariance structure while the loading matrix is restricted to encode module membership. The proposed model can directly accommodate datasets across multiple related classes under the assumption that variables across classes share the same modularity or community structure. While the proposed method is introduced in the context of Gaussian latent variable models, we also demonstrate that it may be extended to latent Bayesian network models. We present experiments on synthetic and fMRI data which demonstrate the capabilities of our approach, in particular showing it successfully scales to high-dimensional data.

References

Airoldi, Edoardo et al. Mixed Membership Stochastic Blockmodels. *J. Mach. Learn. Res.*, 9(2008):1981–2014, 2008. ISSN 1532-4435.

Bertsekas, Dimitri P. *Constrained optimization and Lagrange multiplier methods*. Academic Press, 2014.

Buesing, Lars et al. Clustered factor analysis of multineuronal spike data. *NIPS*, 27:3500–3508, 2014.

Comon, Pierre. Independent component analysis - a new concept? Signal Processing. *Signal Processing*, 36: 287–314, 1994.

Damoiseaux, J S et al. Consistent resting-state networks across healthy subjects. *Proc. Natl. Acad. Sci.*, 103(37): 13848–13853, 2006. ISSN 0027-8424.

Danaher, Patrick et al. The joint graphical lasso for inverse covariance estimation across multiple classes. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 76(2):373–397, 2014.

Dempster, A P. Covariance Selection. *Biometrics*, 28(1): 157, 1972.

Di Martino, Adriana et al. The autism brain imaging data exchange: Towards a large-scale evaluation of the intrinsic brain architecture in autism. *Mol. Psychiatry*, 19(6):659–667, 2014. ISSN 14765578.

Donoho, David and Stodden, Victoria. When does non-negative matrix factorization give a correct decomposition into parts? *NIPS*, pp. 1141–1148, 2004.

Fox, Michael D. and Greicius, Michael. Clinical applications of resting state functional connectivity. *Front. Syst. Neurosci.*, 4, 2010.

Hirayama, Jun Ichiro et al. Characterizing variability of modular brain connectivity with constrained principal component analysis. *PLoS One*, 11(12), 2016.

Hyvärinen, Aapo. Estimation of non-normalized statistical models by score matching. *J. Mach. Learn. Res.*, 6: 695–708, 2005.

Hyvärinen, Aapo. Some extensions of score matching. *Comput. Stat. Data Anal.*, 51(5):2499–2512, 2007.

Keown, Christopher Lee et al. Local functional overconnectivity in posterior brain regions is associated with symptom severity in autism spectrum disorders. *Cell Rep.*, 5(3):567–572, 2013.

Kummerfeld, Erich and Ramsey, Joseph. Causal Clustering for 1-Factor Measurement Models. *KDD*, pp. 1655–1664, 2016.

Ledoit, Olivier and Wolf, Michael. Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *J. Empir. Financ.*, 10(5): 603–621, 2003.

Lin, Lina et al. Estimation of high-dimensional graphical models using regularized score matching. *Electron. J. Stat.*, 10:806–854, 2016.

Marlin, Benjamin and Murphy, Kevin. Sparse Gaussian Graphical Models with Unknown Block Structure. *ICML*, pp. 705–712, 2009.

Monti, Ricardo Pio et al. Learning population and subject-specific brain connectivity networks via Mixed Neighborhood Selection. *Ann. Appl. Stat.*, 11(4):2142–2164, 2017.

Newman, M E J. Modularity and community structure in networks. *Proc. Natl. Acad. Sci.*, 103(23):8577–8582, 2006.

Paatero, Pentti and Tapper, Unto. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 5(2):111–126, 1994.

Pearl, Judea. *Causality*. Cambridge University Press, 2009.

Sasaki, Hiroaki et al. Simultaneous estimation of non-gaussian components and their correlation structure. *Neural Comput.*, 29:2887–2924, 2017.

Seung, H. Sebastian and Lee, Daniel D. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

Shimizu, Shohei et al. A Linear Non-Gaussian Acyclic Model for Causal Discovery. *J. Mach. Learn. Res.*, 7: 2003–2030, 2006.

Shimizu, Shohei et al. Estimation of linear non-Gaussian acyclic models for latent factors. *Neurocomputing*, 72 (7-9):2024–2027, 2009.

Sigg, Christian D and Buhmann, Joachim M. Expectation-maximization for sparse and non-negative PCA. *ICML*, pp. 960–967, 2008.

Silva, Ricardo et al. Learning the structure of linear latent variable models. *J. Mach. Learn. Res.*, 7:191–246, 2006.

Sporns, Olaf and Betzel, Richard F. Modular Brain Networks. *Annu. Rev. Psychol.*, 67(1):613–640, 2016.

Tipping, M. E. and Bishop, C. M. Probablistic Principle Component Analysis. *J. R. Stat. Soc. Ser. B (Statistical Methodol.*, 11(2):150–210, 1999.

Varoquaux, Gaël et al. Brain covariance selection: better individual functional connectivity models using population prior. *NIPS*, 2010.

Supplementary Material

Maximum likelihood estimation

In this supplement we derive the maximum likelihood estimation algorithm for the proposed Gaussian factor analysis model. We note that the log-likelihood associated with the proposed model is:

$$\mathcal{L} = \sum_{i=1}^{N} p \log 2\pi + \log \det \Sigma^{(i)} + \text{tr}\left(\Sigma^{(i)^{-1}} K^{(i)}\right). \tag{10}$$

In the case of the loading matrix, the gradient update is defined as:

$$\frac{\partial \mathcal{L}}{\partial W} = \sum_{i=1}^{N} \frac{\partial \mathcal{L}}{\partial \Sigma^{(i)}} \frac{\partial \Sigma^{(i)}}{\partial W} = \sum_{i=1}^{N} \left(\underbrace{-\Sigma^{(i)^{-1}} + \Sigma^{(i)^{-1}} K^{(i)} \Sigma^{(i)^{-1}}}_{M^{(i)}} \right) W G^{(i)} \tag{11}$$

We note that the main computational burden is associated with computing $M^{(i)}$. Using the Sherman-Woodbury identity, we may write $M^{(i)}$ as:

$$\begin{aligned} M^{(i)} = &- v^{(i)^{-1}} I + v^{(i)^{-1}} W A^{(i)} W^T + v^{(i)^{-2}} K^{(i)} \\ &- 2 v^{(i)^{-2}} W A^{(i)} W^T K^{(i)} \\ &+ v^{(i)^{-2}} W A^{(i)} W^T K^{(i)} W A^{(i)} W^T, \end{aligned}$$

from which it follows that computing the gradient of the log-likelihood with respect to the loading matrix, W, incurs a computational cost of $\mathcal{O}(p^3)$.

In the case of the latent connectivity matrix, $G^{(i)}$, the update is equivalent to the score matching algorithm. This follows from:

$$\frac{\partial \mathcal{L}}{\partial G^{(i)}} = \sum_{i=1}^{N} \frac{\partial \mathcal{L}}{\partial \Sigma^{(i)}} \frac{\partial \Sigma^{(i)}}{\partial G^{(i)}} \tag{12}$$

$$= \sum_{i=1}^{N} \left(-\Sigma^{(i)^{-1}} + \Sigma^{(i)^{-1}} K^{(i)} \Sigma^{(i)^{-1}} \right) W^T W. \tag{13}$$

Setting equation (13) to equal zero implies that $I = K^{(i)} \Sigma^{(i)^{-1}}$, which after re-arranging yields:

$$G^{(i)} = W^T K^{(i)} W - v^{(i)} I. \tag{14}$$

Score matching estimation

In this supplement we provide a detailed derivation for the score matching algorithm presented in Section 4. We begin by explicitly writing the score matching objective in terms of parameters W, $\{G^{(i)}\}$ and $\{v^{(i)}\}$. This is specified as:

$$J = \sum_{i=1}^{N} -\text{tr}\left(\Omega^{(i)}\right) + \frac{1}{2}\text{tr}\left(\Omega^{(i)} \Omega^{(s)} K^{(i)}\right) \tag{15}$$

$$= \sum_{i=1}^{N} \Big[-v^{(i)^{-1}} \text{tr}\,(I) + v^{(i)^{-1}} \text{tr}(A^{(i)}) + \frac{1}{2} v^{(i)^{-2}} \text{tr}(K^{(i)}) \tag{16}$$

$$-v^{(i)^{-2}} \text{tr}(W^T K^{(i)} W A^{(i)}) + \frac{1}{2} v^{(i)^{-2}} \text{tr}(W^T K^{(i)} W A^{(i)} A^{(i)}) \Big] \tag{17}$$

where $A^{(i)} = G^{(i)}(G^{(i)} + v^{(i)}I)^{-1}$ as in the original text. We may now directly compute the derivatives with respect to each of the parameters in the proposed latent variable model. The derivative for the loading matrix is:

$$\frac{\partial J}{\partial W} = \sum_{i=1}^{N} {v^{(i)}}^{-2} K^{(i)} W \left(\frac{1}{2} A^{(i)} A^{(i)} - A^{(i)} \right). \tag{18}$$

The derivative with respect to latent connectivities can be obtained via the chain rule as:

$$\begin{aligned}
\frac{\partial J}{\partial G^{(i)}} &= \frac{\partial J}{\partial A^{(i)}} \frac{\partial A^{(i)}}{\partial G^{(i)}} && (19)\\
&= {v^{(i)}}^{-2} \left[v^{(i)} I - W^T K^{(i)} W \left(I - A^{(i)} \right) \right] \frac{\partial A^{(i)}}{\partial G^{(i)}} && (20)\\
&= {v^{(i)}}^{-2} \left[v^{(i)} I - W^T K^{(i)} W \left(I - A^{(i)} \right) \right] \left[(G^{(i)} + v^{(i)} I)^{-1} \left(I - A^{(i)} \right) \right]. && (21)
\end{aligned}$$

We note that setting equation (21) to zero implies that the middle term must be zero, as both $(G^{(i)} + v^{(i)}I)^{-1}$ and $(I - A^{(i)})$ cannot be zero. By equating the middle term with zero, we obtain:

$$v^{(i)} \left(W^T K^{(i)} W \right)^{-1} = I - A^{(i)}, \tag{22}$$

which after re-arranging yields:

$$A^{(i)} = G^{(i)}(G^{(i)} + v^{(i)}I)^{-1} = I - v^{(i)} \left(W^T K^{(i)} W \right)^{-1}. \tag{23}$$

Finally, we may re-arranging for $G^{(i)}$ to obtain:

$$G^{(i)} = W^T K^{(i)} W - v^{(i)} I \tag{24}$$

which is the same update as obtained in equation (14) above. Before deriving the derivative of the score matching objective with respect to $v^{(i)}$, we state the following identities which will of use later on:

- We may eigendecompose the latent connectivity $G^{(i)}$ as follows:

  $$G^{(i)} = V_i D_i V_i^T, \tag{25}$$

  where V_i is a matrix of eigenvectors and D_i is a diagonal matrix of eigenvalues $d_1, \ldots, d_k$. Therefore we may write $A^{(i)}$ as follows:

  $$\begin{aligned}
  A^{(i)} &= G^{(i)}(G^{(i)} + v^{(i)}I)^{-1} \\
  &= V_i \; D_i \; V_i^T \; V_i \; \text{diag} \left(\frac{1}{d_j + v^{(i)}} \right) \; V_i^T \\
  &= V_i \; \text{diag} \left(\frac{d_j}{d_j + v^{(i)}} \right) \; V_i^T
  \end{aligned}$$

  As a result, we can compute the derivative of $A^{(i)}$ with respect to $v^{(i)}$ as follows:

  $$\frac{\partial A^{(i)}}{\partial v^{(i)}} = V_i \; \text{diag} \left(\frac{-d_j}{(d_j + v^{(i)})^2} \right) \; V_i^T = \tilde{D}_i. \tag{26}$$

  Furthermore, because V_i are eigenvectors, we have that $\frac{\partial \text{tr}(A^{(i)})}{\partial v^{(i)}} = \text{tr}(\tilde{D}_i)$.

- Using the same arguments, we may write the derivative of $A^{(i)}A^{(i)}$ with respect to $v^{(i)}$ as follows:

  $$\frac{\partial A^{(i)} A^{(i)}}{\partial v^{(i)}} = V_i^T \; \text{diag} \left(\frac{-2 d_j^2}{(d_j + v^{(i)})^3} \right) \; V_i^T = \tilde{\tilde{D}}_i \tag{27}$$

Using equations (26) and (27) we may therefore write the derivative of the score matching objective with respect to $v^{(i)}$ as follows:

$$\frac{\partial J}{\partial v^{(i)}} = {v^{(i)}}^{-3}\text{tr}\left(v^{(i)}I - K^{(i)}\right) \tag{28}$$

$$+ {v^{(i)}}^{-3}\text{tr}\left(W^T K^{(i)} W\left[2A^{(i)} - v^{(i)}\tilde{D}_i - A^{(i)}A^{(i)} + v^{(i)}\tilde{\tilde{D}}_i\right]\right) \tag{29}$$

$$- {v^{(i)}}^{-2}\text{tr}\left(A^{(i)} - v^{(i)}\tilde{D}_i\right) \tag{30}$$

As such, we define:

$$H_1^{(i)} = 2A^{(i)} - v^{(i)}\tilde{D}_i - A^{(i)}A^{(i)} + v^{(i)}\tilde{\tilde{D}}_i$$
$$H_2^{(i)} = -{v^{(i)}}^{-2}\text{tr}\left(A^{(i)} - v^{(i)}\tilde{D}_i\right)$$

Additional experiments: cluster recovery

Results in Section 6 demonstrated that the proposed method is able to reliably recover the loading matrix, W, in terms of mean squared error. In this section we provide additional results demonstrating that the clusters inferred from the estimated loading matrix accurately reflect the true clustering of variables. The adjusted Rand Index was employed in order to quantify the similarity between the estimated and true clusterings. Results are shown in Figure 6 for Gaussian factor analysis models and latent Bayesian networks along the top and bottom row respectively. In each case, we note that the proposed model is able to accurately cluster variables, inclusively when compared with traditional clustering algorithms such as k-means and hierarchical clustering. Furthermore, we note that as the number of classes increases from $N = 1$ to $N = 10$, the accuracy of the proposed method improves as it is able to combine observations across classes.

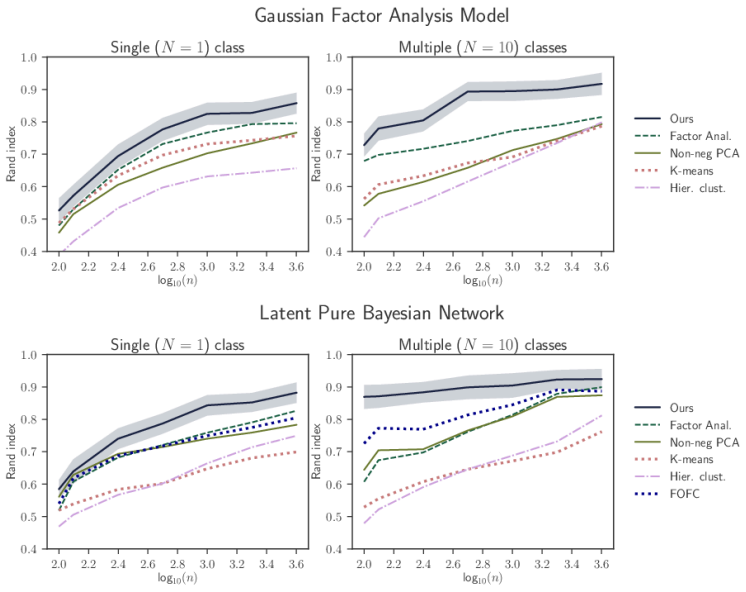

Figure 6: Adjusted Rand index scores for variable clustering as inferred by the estimated loading matrices. Results are shown for Gaussian factor analysis models (top) and latent Bayesian networks (bottom) as well as for $N = 1$ and $N = 10$ classes in the left and right columns respectively. Shaded regions correspond to 95% error bars.